# BST 676 – Spring 2010 – Dr. Charnigo

## Unit II: Techniques for Point Estimation

### a. Motivating Case Study #1: Obtaining degrees of freedom for a two-sample T test

Suppose that we wish to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ at significance level $\alpha \in (0, 1)$, where $\mu_1$ and $\mu_2$ are means from two normal populations. An introductory methods course such as STA 580 teaches that the appropriate procedure depends on whether we are willing to assume equality of the population variances, $\sigma_1^2 = \sigma_2^2$.

If we are willing to assume equality of the population variances, then with obvious notation we define the test statistic

$$t_{equal} := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{pooled}(1/n_1 + 1/n_2)}} \quad \text{with} \quad s^2_{pooled} := \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and reject the null hypothesis if $|t_{equal}|$ exceeds $t_{n_1+n_2-2,1-\alpha/2}$, the $1-\alpha/2$ quantile of the T distribution on $n_1 + n_2 - 2$ degrees of freedom. (If $\psi_{n_1+n_2-2}$ denotes the cumulative distribution function for the T distribution on $n_1 + n_2 - 2$ degrees of freedom, then $t_{n_1+n_2-2,1-\alpha/2}$ satisfies $\psi_{n_1+n_2-2}(t_{n_1+n_2-2,1-\alpha/2}) = 1 - \alpha/2$.)

If we are unwilling to assume equality of the population variances, then with obvious notation we define the test statistic

$$t_{unequal} := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

and reject the null hypothesis if $|t_{unequal}|$ exceeds $t_{df,1-\alpha/2}$, where

$$df := \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}. \tag{1}$$

The result of equation (1) may be floored (i.e., rounded down to the next lower integer) if we are relying on a table in the back of a methods textbook for critical values. However, in mathematical principle, the degrees of freedom parameter that indexes the family of T distributions can be any positive real number, not just any positive integer. In fact, computer programs like SAS and R can perform computations involving T distributions on non-integer degrees of freedom.

Students in STA 580 dread using equation (1) to calculate degrees of freedom because, typically, both the numerator and denominator are very small numbers. This leads to calculator displays in scientific notation and also the potential for considerable roundoff error. (What usually determines the magnitude of roundoff error is the number of nonzero digits that are retained, not the number of decimal places. So, for example, rounding 0.000044 to 0.00004 introduces a big error. Unfortunately, many people tend to think in terms of retaining a certain number of decimal places.)

We can obtain some insight from (1) by considering two extreme cases. First, suppose that $s_1^2 = s_2^2 =: s^2$ and $n_1 = n_2 =: n$. Then we have

$$\begin{aligned} df &= \frac{(s^2/n + s^2/n)^2}{(s^2/n)^2/(n-1) + (s^2/n)^2/(n-1)} \\ &= \frac{(1/n + 1/n)^2}{(1/n)^2/(n-1) + (1/n)^2/(n-1)} \\ &= \frac{4}{1/(n-1) + 1/(n-1)} \\ &= \frac{4(n-1)}{2} \\ &= n_1 + n_2 - 2. \end{aligned} \tag{2}$$

Second, suppose that $s_2^2/n_2$ is negligibly small compared to $s_1^2/n_1$. Then we have

$$df \approx \frac{(s_1^2/n_1)^2}{(s_1^2/n_1)^2/(n_1-1)} = (n_1 - 1). \tag{3}$$

Results (2) and (3) suggest that $df$ will be close to $n_1 + n_2 - 2$ when the sample variances and sample sizes are almost equal, while $df$ will be much less than $n_1 + n_2 - 2$ (and perhaps only slightly larger than $n_1 - 1$ or $n_2 - 1$) when the sample variances or sample sizes are grossly unequal. Paradoxically, you can actually lose degrees of freedom by increasing your sample sizes! Work the following out at home: Supposing that $s_1^2 = s_2^2 = 1$, is $df$ larger when $n_1 = n_2 = 40$ or when $n_1 = 60$ and $n_2 = 600$?

In any event, we still do not know where (1) comes from.

We will revisit this case study at the end of Unit II.

**b. Motivating Case Study #2: Should cancer rate estimates be zero when there are no cases?**

Consider the following fictional data set. The entries in the "Cases" column represent the number of people in a geographic region who developed a certain rare form of cancer during 2009, while the entries in the "Population" column represent the number of people living in that geographic region.

| Region | Cases | Population | Cases per 100,000 Population |
|---:|---:|---:|---:|
| A | 1 | 50,000 | 2.0 |
| B | 0 | 80,000 | 0.0 |
| C | 6 | 120,000 | 5.0 |
| D | 5 | 200,000 | 2.5 |
| E | 2 | 250,000 | 0.8 |
| F | 6 | 300,000 | 2.0 |

If our interest lies in describing what actually happened in 2009, then the observed incidence rate for region B is 0.

However, suppose that we regard each observed incidence rate as a realization of a random variable and that our real interest lies in estimating the expected value of this random variable. The rationale for desiring to estimate such an expected value is that it is simultaneously interpretable as: (i) our best prediction for that region's observed incidence rate in 2010; and, (ii) the risk that a randomly selected person from that region becomes a case in 2010.

The second interpretation especially makes clear that 0, despite being the observed incidence rate for region B in 2009, is not a reasonable estimate of the expected value.

**Discussion Question**. Prior to studying Unit II, can you think of any way to validly obtain a nonzero estimate of the risk for a randomly selected person from region B in 2010?

We will revisit this case study at the end of Unit II.

**c. Method of moments (Cf. pp. 357-363 of Larsen and Marx)**

A typical situation in parametric statistical inference is as follows. An observed data set is regarded as the realization of a random sample from a probability mass function or a probability density function that is specified up to an unknown parameter. (The parameter may be a vector.) We want to use the observed data set to estimate that parameter.

Example #1. Suppose that

$$Y_1, Y_2, Y_3, Y_4 \overset{iid}{\sim} f_Y(y;\theta) := \theta \exp[-y\theta] 1_{\{y>0\}},$$

where $\theta \in \Theta := (0, \infty)$ is an unknown parameter. We refer to $\Theta$ as the "parameter space". In general, the parameter space represents all mathematically possible values for $\theta$ or a subset of such values chosen by the data analyst. Regarding the notation, "iid" is shorthand for independent and identically distributed. The expression $f_Y(y;\theta)$ is the probability density function common to $Y_1, Y_2, Y_3, Y_4$ evaluated at the number $y$. (Can you give the name of this probability density function?) To emphasize that $f_Y(y;\theta)$ depends on $\theta$, we include $\theta$ in the notation. Now suppose that we observe $Y_1 = 2.4$, $Y_2 = 3.8$, $Y_3 = 5.7$, and $Y_4 = 10.2$. What is our best guess for $\theta$? We will return to this example later.

Example #2. Suppose that

$$Y_1, Y_2, Y_3, Y_4 \overset{iid}{\sim} f_Y(y;\theta) := \exp[-\theta]\theta^y/y!\ 1_{\{y\in\{0,1,2,\ldots\}\}},$$

where $\theta \in \Theta := (0, \infty)$ is an unknown parameter. The expression $f_Y(y;\theta)$ is the probability mass function common to $Y_1, Y_2, Y_3, Y_4$ evaluated at the number $y$. (Can you give the name of this probability mass function?) Now suppose that we observe $Y_1 = 1$, $Y_2 = 5$, $Y_3 = 8$, and $Y_4 = 3$. What is our best guess for $\theta$?

Example #3. Suppose that

$$Y_1, Y_2, Y_3, Y_4 \overset{iid}{\sim} f_Y(y;\theta) := (2\pi\theta_2)^{-1/2} \exp[-(y-\theta_1)^2/(2\theta_2)],$$

where $\theta = (\theta_1, \theta_2)^T \in \Theta := (-\infty, \infty) \times (0, \infty)$ is an unknown vector parameter. The first component of $\theta$ is the mean, while the second component of $\theta$ is the variance. Above, "T" stands for transpose and is used here since ordinarily a vector is taken to be a column rather than a row. Explicitly, the transpose of the row $(\theta_1, \theta_2)$ is the column $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$. Now suppose that we observe $Y_1 = 1.5$, $Y_2 = 0.8$, $Y_3 = 3.6$, and $Y_4 = 1.7$. What is our best guess for $\theta$?

To address the three examples above, we will employ the following strategy. Suppose that $\theta$ is a vector of dimension $s$ and that $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} f_Y(y; \theta)$, a probability mass function or probability density function that has finite first through $s^{th}$ moments, which we denote by $m_1(\theta)$ through $m_s(\theta)$.

We have

$$E_\theta[n^{-1}\sum_{i=1}^n Y_i] = m_1(\theta), \quad E_\theta[n^{-1}\sum_{i=1}^n Y_i^2] = m_2(\theta), \quad \cdots, \quad E_\theta[n^{-1}\sum_{i=1}^n Y_i^s] = m_s(\theta).$$

Above, the $\theta$ subscript on the expectation operator emphasizes that the expected values depend on $\theta$. Moreover, by the Weak Law of Large Numbers, we have

$$n^{-1}\sum_{i=1}^n Y_i \xrightarrow{P} m_1(\theta), \quad n^{-1}\sum_{i=1}^n Y_i^2 \xrightarrow{P} m_2(\theta), \quad \cdots, \quad n^{-1}\sum_{i=1}^n Y_i^s \xrightarrow{P} m_s(\theta)$$

as $n \to \infty$.

The preceding considerations suggest that we can reasonably estimate $\theta$ by solving the equations

$$n^{-1}\sum_{i=1}^n Y_i = m_1(\widehat{\theta}), \quad n^{-1}\sum_{i=1}^n Y_i^2 = m_2(\widehat{\theta}), \quad \cdots, \quad n^{-1}\sum_{i=1}^n Y_i^s = m_s(\widehat{\theta}) \tag{4}$$

for $\widehat{\theta}$. The $\widehat{\theta}$ so determined is called an "estimator" of $\theta$, while the numerical value realized by $\widehat{\theta}$ for a particular observed data set is called an "estimate".

The strategy of employing (4) is called the "method of moments".

In the special case that $s = 2$, we often replace (4) by

$$n^{-1}\sum_{i=1}^n Y_i = m_1(\widehat{\theta}), \quad (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 = v(\widehat{\theta}), \tag{5}$$

where $v(\theta)$ denotes the variance associated with the probability mass function or probability density function $f_Y(y; \theta)$.

An important caveat to the method of moments is that in some cases there may not exist a solution to (4) or (5). This is illustrated in a fourth example below.

Example #1, continued. We have $m_1(\theta) = 1/\theta$, so we obtain $\widehat{\theta}$ by solving

$$n^{-1} \sum_{i=1}^{n} Y_i = 1/\widehat{\theta}.$$

The method of moments estimator of $\theta$ is thus

$$\widehat{\theta} = n / \sum_{i=1}^{n} Y_i.$$

With $Y_1 = 2.4$, $Y_2 = 3.8$, $Y_3 = 5.7$, and $Y_4 = 10.2$ we have $\sum_{i=1}^{n} Y_i = 22.1$ and so the method of moments estimate of $\theta$ is $\widehat{\theta} = 4/22.1 = 0.181$.

Example #2, continued. We have $m_1(\theta) = \theta$, so

$$\widehat{\theta} = n^{-1} \sum_{i=1}^{n} Y_i.$$

With $Y_1 = 1$, $Y_2 = 5$, $Y_3 = 8$, and $Y_4 = 3$ we have $\widehat{\theta} = 4.25$.

Example #3, continued. We have $m_1(\theta) = \theta_1$ and $v(\theta) = \theta_2$. This yields

$$\widehat{\theta} =$$

With $Y_1 = 1.5$, $Y_2 = 0.8$, $Y_3 = 3.6$, and $Y_4 = 1.7$ we have

$$\widehat{\theta} =$$

Example #4. Consider the probability density function

$$f_Y(y; \theta) := (1 - \theta_1)(2\pi)^{-1/2} \exp[-(y + \theta_2)^2/2] + \theta_1 (2\pi)^{-1/2} \exp[-(y - \theta_2)^2/2],$$

where $\theta = (\theta_1, \theta_2)^T \in \Theta := [0, 1] \times (-\infty, \infty)$. One can show (do so at home, for practice) that

$$m_1(\theta) = \theta_2(2\theta_1 - 1), \quad v(\theta) = 1 + 4\theta_2^2 \theta_1 (1 - \theta_1) \geq 1.$$

Hence, there is no solution to

$$m_1(\widehat{\theta}) = n^{-1} \sum_{i=1}^{n} Y_i, \quad v(\widehat{\theta}) = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

and thus no method of moments estimate, if we should happen to have $(n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 < 1$.

**d. Maximum likelihood (Cf. pp. 347-357 of Larsen and Marx)**

Another approach to parameter estimation entails identifying the element of the parameter space for which the probability of observing the present data set is maximized.

Example #2, continued. Let $\zeta$ be a generic element of the parameter space $\Theta := (0, \infty)$. If we had $\theta = \zeta$, then for generic nonnegative integers $y_1, y_2, y_3, y_4$ we would have

$$P_{\theta=\zeta}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) = \frac{\exp[-4\zeta]\zeta^{y_1+y_2+y_3+y_4}}{y_1!\ y_2!\ y_3!\ y_4!}. \tag{6}$$

On the one hand, expression (6) can be viewed as a function of $y_1, y_2, y_3, y_4$. Indeed, we call it the joint probability mass function of $Y_1, Y_2, Y_3, Y_4$. On the other hand, expression (6) can also be viewed as a function of $\zeta$. In that case, we call it the likelihood function and use the notation $L(\zeta; \mathbf{y})$, where the letter $L$ recalls the word "likelihood" and $\mathbf{y}$ is a shorthand for $y_1, y_2, y_3, y_4$.

To better understand the role of the likelihood function in estimating $\theta$, suppose that we have observed $Y_1 = 1$, $Y_2 = 5$, $Y_3 = 8$, and $Y_4 = 3$. The probability of such an observation if $\theta = 1$ is obtained by substituting 1 for $\zeta$ in (6),

$$\frac{\exp[-4]}{1!\ 5!\ 8!\ 3!} = 6.31 \times 10^{-10},$$

while the probability of such an observation if $\theta = 4$ is obtained by substituting 4 for $\zeta$ in (6),

$$\frac{\exp[-16]4^{1+5+8+3}}{1!\ 5!\ 8!\ 3!} = 6.66 \times 10^{-5}.$$

Thus, observing $Y_1 = 1$, $Y_2 = 5$, $Y_3 = 8$, and $Y_4 = 3$ is about $100,000$ times more likely if $\theta = 4$ than if $\theta = 1$. In other words, $\theta = 4$ is more consistent with the observed data than $\theta = 1$. So, guessing that $\theta = 4$ is more reasonable than guessing that $\theta = 1$.

We can go one step further and maximize (6) with respect to $\zeta \in \Theta$. Since maximizing (6) is equivalent to maximizing its natural logarithm, we may as well work with the "log likelihood"

$$l(\zeta; \mathbf{y}) := \log L(\zeta; \mathbf{y}) = -4\zeta + (y_1 + y_2 + y_3 + y_4) \log \zeta + C(\mathbf{y}), \tag{7}$$

where $C(\mathbf{y})$ is a term that does not depend on $\zeta$. In this example, we have $C(\mathbf{y}) = -(\log y_1! + \log y_2! + \log y_3! + \log y_4!)$.

Suppose that $y_1 + y_2 + y_3 + y_4 > 0$. Differentiating the log likelihood in $\zeta$ yields

$$\frac{\partial l(\zeta; \mathbf{y})}{\partial \zeta} = -4 + (y_1 + y_2 + y_3 + y_4)/\zeta,$$

which equals 0 if and only if $\zeta = (y_1 + y_2 + y_3 + y_4)/4 =: \widehat{\theta}$. We know that a derivative equaling 0 does not guarantee a local maximum, much less a global maximum. However, in this case we obtain a global maximum of $l(\zeta; \mathbf{y})$ because its derivative is obviously positive when $\zeta \in (0, \widehat{\theta})$ and negative when $\zeta \in (\widehat{\theta}, \infty)$. That is, $\widehat{\theta}$ maximizes both the log likelihood and the likelihood. We refer to $\widehat{\theta}$ as a maximum likelihood estimate of $\theta$.

Remark #1. In the above example, the method of moments and maximum likelihood yield the same estimate. That is not always true.

Remark #2. In general, there is no guarantee that a maximum likelihood estimate exists or that, if it does, it is unique. Suppose that we had $y_1 = y_2 = y_3 = y_4 = 0$ in the above example. Then the derivative of the log likelihood would be negative for all $\zeta \in (0, \infty)$. This would seem to suggest that we should take $\widehat{\theta} := 0$, but unfortunately 0 is not in the parameter space. One way to fix that problem is to artificially expand the parameter space to include 0. (In that case, we interpret "Poisson with parameter 0" as representing a degenerate random variable that assumes the value 0 with probability 1.) Also, if we have a unique maximizer of the likelihood, then we may refer to it as "the" maximum likelihood estimate rather than as "a" maximum likelihood estimate.

Remark #3. The number $(y_1 + y_2 + y_3 + y_4)/4$ is called the maximum likelihood estimate, while the random variable $(Y_1 + Y_2 + Y_3 + Y_4)/4$ is called the maximum likelihood estimator. Statisticians routinely use $\widehat{\theta}$ to represent both the estimate and the estimator.

Remark #4. An even more substantial notational problem is that statisticians often express the likelihood as a function of $\theta$ rather than as a function of $\zeta$. This unfortunate convention obscures the distinction between the actual parameter that we are trying to estimate and a generic element of the parameter space.

Example #1, continued. If we try to repeat the developments of the last two pages with the probability density function $f_Y(y;\theta) := \theta \exp[-y\theta] 1_{\{y>0\}}$, we run into the immediate problem that

$$P_{\theta=\zeta}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) = 0$$

for any positive real numbers $y_1, y_2, y_3, y_4$. (Why?)

What statisticians recommend for continuous $Y_1, Y_2, Y_3, Y_4$ is that the likelihood function be defined using the joint probability density function. In this example, we have

$$L(\zeta; \mathbf{y}) := \zeta^4 \exp[-\zeta(y_1 + y_2 + y_3 + y_4)]. \tag{8}$$

The rationale for using (8) is the following limiting argument.

Let $\epsilon$ be a small positive number. For any positive real number $y_1$, we have

$$P_{\theta=\zeta}(Y_1 \in [y_1, y_1 + \epsilon]) = \int_{y_1}^{y_1+\epsilon} \zeta \exp[-\zeta t] dt \approx \epsilon \zeta \exp[-\zeta y_1],$$

the approximation becoming better in the limit as $\epsilon$ approaches 0. Likewise, we have

$$\begin{aligned} & P_{\theta=\zeta}(Y_1 \in [y_1, y_1+\epsilon], Y_2 \in [y_2, y_2+\epsilon], Y_3 \in [y_3, y_3+\epsilon], Y_4 \in [y_4, y_4+\epsilon]) \quad (9) \\ \approx\ & \epsilon^4 \zeta^4 \exp[-\zeta(y_1 + y_2 + y_3 + y_4)] \\ =\ & \epsilon^4 L(\zeta; \mathbf{y}). \end{aligned}$$

Maximizing (8) with respect to $\zeta \in \Theta$ is approximately the same as maximizing (9), the approximation becoming better in the limit as $\epsilon$ approaches 0. Moreover, since maximizing (9) is reasonable, so is maximizing (8).

Once we have defined the likelihood, the rest of this example is easy. The log likelihood is

$$l(\zeta; \mathbf{y}) := \log L(\zeta; \mathbf{y}) = 4 \log \zeta - \zeta(y_1 + y_2 + y_3 + y_4),$$

and its derivative is

$$\frac{\partial l(\zeta; \mathbf{y})}{\partial \zeta} = 4/\zeta - (y_1 + y_2 + y_3 + y_4),$$

which equals 0 if and only if $\zeta = \widehat{\theta} :=$ We can readily verify that $\widehat{\theta}$ is indeed a global maximizer (do so at home, for practice).

Example #5. A common difficulty faced by students is that they see the ease with which setting to 0 the derivative of $l(\zeta; \mathbf{y})$ yields a maximum likelihood estimate in many cases and then assume that this strategy will always work. The present example illustrates a situation in which this strategy fails and clarifies what can be done in such a situation.

Consider the probability density function $f_Y(y; \theta) := 1/\theta$ for $y \in [0, \theta]$, where $\theta \in \Theta := (0, \infty)$. Suppose that $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} f_Y(y; \theta)$. Given observed values $y_1, y_2, \ldots, y_n$, the likelihood and log likelihood would appear to be

$$\frac{1}{\zeta^n} \tag{10}$$

and

$$-n \log \zeta. \tag{11}$$

The derivative of (11) is $-n/\zeta$, which cannot be made equal to 0. At this point, the student is stuck.

Our problem is that (10) ignores the requirement of $f_Y(y; \theta)$ that $y \in [0, \theta]$. Moreover, a moment's reflection reveals that maximizing (10) does not even make sense since (10) does not depend on the data!

To overcome our problem, we note that, if $Y_1$ has been observed to equal $y_1$, then $\theta$ must be greater than or equal to $y_1$. Likewise, if $Y_1, Y_2, \ldots, Y_n$ have been observed to equal $y_1, y_2, \ldots, y_n$, then $\theta$ must be greater than or equal to the maximum of $y_1, y_2, \ldots, y_n$. This yields the correct definition of the likelihood,

$$L(\zeta; \mathbf{y}) := \frac{1}{\zeta^n} 1_{\{\zeta \geq \max\{y_1, y_2, \ldots, y_n\}\}}. \tag{12}$$

Since $1/\zeta^n$ is strictly decreasing in $\zeta \in (0, \infty)$, expression (12) is maximized at the smallest value of $\zeta$ for which the indicator does not vanish. This is $\max\{y_1, y_2, \ldots, y_n\}$. In summary, the maximum likelihood estimate in this example is $\widehat{\theta} := \max\{y_1, y_2, \ldots, y_n\}$.

As a general rule of thumb, reasoning like that in the present example rather than differentiation of the log likelihood will be appropriate whenever $\{y \in \mathbb{R} : f_Y(y; \theta) > 0\}$ depends on $\theta$. We refer to $\{y \in \mathbb{R} : f_Y(y; \theta) > 0\}$ as the "support set" of $f_Y(y; \theta)$. In the present example, the support set is $[0, \theta]$, which depends on $\theta$. In Example #1, the support set was $(0, \infty)$, which does not depend on $\theta$.

**e. Bayesian posterior mode and posterior mean (Cf. pp. 410-423 of Larsen and Marx)**

Until now we have maintained a frequentist perspective, which regards $\theta$ as an unknown but fixed number. In contrast, a Bayesian perspective regards $\theta$ as a random variable. This random variable is described by a prior distribution before we observe the data and by a posterior distribution after we observe the data. Even if we do not wish to abandon a frequentist perspective, we can still go through the mechanics of specifying a prior distribution, determining a posterior distribution, and then defining $\widehat{\theta}$ to be either the mean or the mode of the posterior distribution.

Example #1, continued. Recall that we put

$$f_Y(y;\theta) := \theta \exp[-y\theta]$$

for $y \in (0, \infty)$. Suppose that, before the data are observed, $\theta$ has the probability density function

$$p(\theta) := C(a,b)\theta^{a-1} \exp[-b\theta]$$

for $\theta \in \Theta := (0, \infty)$, where $a$ and $b$ are specified positive constants and $C(a, b)$ is chosen so that

$$\int_0^\infty p(\theta)\ d\theta = 1.$$

In fact, we recognize that $p(\theta)$ is a gamma probability density function with shape parameter $a$ and rate parameter $b$, so

$$C(a,b) = \frac{b^a}{\Gamma[a]},$$

where

$$\Gamma[x] := \int_0^\infty t^{x-1} \exp[-t]\ dt$$

is the gamma function with domain $(0, \infty)$.

Suppose that we observe $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, where $y_1, y_2, \ldots, y_n$ are positive real numbers. What is the posterior distribution of $\theta$?

We may recall Bayes' rule from BST 675. For two events $A$ and $B$, we have

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(\bar{B})P(A|\bar{B})},$$

as long as this expression is well defined (in particular, does not entail division by zero). Moreover, if $B_1, B_2, \ldots, B_m$ are mutually exclusive and collectively exhaustive, we have

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{\sum_{k=1}^{m} P(B_k)P(A|B_k)}. \tag{13}$$

By analogy to (13), we might wish to write

$$P(\theta = \zeta | Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \frac{P(\theta = \zeta)L(\zeta; \mathbf{y})}{\sum_{\eta \in \Theta} P(\theta = \eta)L(\eta; \mathbf{y})}, \tag{14}$$

where $\zeta$ is a generic element of $\Theta$. However, formula (14) will not work unless the prior distribution of $\theta$ is discrete, which is not a reasonable specification when $\Theta$ is an interval. The appropriate formula, when the prior distribution of $\theta$ is continuous, is

$$p(\theta; \mathbf{y}) = \frac{p(\theta)L(\theta; \mathbf{y})}{\int_{\eta \in \Theta} p(\eta)L(\eta; \mathbf{y})\, d\eta}, \tag{15}$$

where $p(\theta; \mathbf{y})$ denotes the probability density function for $\theta$ after the data are observed. Note that the denominator of (15) is a function of $\mathbf{y}$ but not of $\theta$ or of $\eta$, which is integrated out. The implication is that, if we can recognize $p(\theta)L(\theta; \mathbf{y})$ as the "kernel" of a familiar probability density function, we don't need to explicitly evaluate $\int_{\eta \in \Theta} p(\eta)L(\eta; \mathbf{y})\, d\eta$.

Continuing with our example, we have

$$\begin{aligned} p(\theta; \mathbf{y}) &\propto p(\theta)L(\theta; \mathbf{y}) \\ &\propto \theta^{a-1} \exp[-b\theta]\theta^{n} \exp[-(y_1 + y_2 + \cdots + y_n)\theta] \\ &= \theta^{a+n-1} \exp[-(b + n\bar{y})\theta], \end{aligned} \tag{16}$$

where the symbol $\propto$ (read "is proportional to") allows us to discard multiplicative factors, such as $C(a, b)$ and $\int_{\eta \in \Theta} p(\eta)L(\eta; \mathbf{y})\, d\eta$, that do not depend on $\theta$. In fact, we see from (16) that the posterior distribution of $\theta$ is gamma with shape parameter $a + n$ and rate parameter $b + n\bar{y}$. If we wanted to do so, we could write out that

$$p(\theta; \mathbf{y}) = \frac{(b + n\bar{y})^{a+n}}{\Gamma[a + n]} \theta^{a+n-1} \exp[-(b + n\bar{y})\theta]$$

for $\theta \in (0, \infty)$. Note that this did not require us to explicitly evaluate $\int_{\eta \in \Theta} p(\eta)L(\eta; \mathbf{y})\, d\eta$.

If we are really frequentists and want to reduce the posterior distribution (16) into a representative single number, then we can employ either the posterior mode or the posterior mean.

The posterior mode is defined as the maximizer of $p(\theta; \mathbf{y})$ over $\theta \in \Theta$. For scalar $\theta$, we can look for the posterior mode by setting

$$\frac{\partial p(\theta; \mathbf{y})}{\partial \theta} = 0,$$

subject to the usual caveat that setting a derivative equal to zero does not always yield a local maximum, much less a global maximum. In the present example (verify this at home, for practice), the posterior mode is

$$\frac{a+n-1}{b+n\bar{y}}.$$

The posterior mean is defined as

$$\int_0^\infty \theta \ p(\theta; \mathbf{y}) \ d\theta,$$

presuming that the integral is absolutely convergent. In the present example, the posterior mean is $(a+n)/(b+n\bar{y})$ because the mean of a gamma random variable with shape parameter $\alpha$ and rate parameter $\beta$ is $\alpha/\beta$. If we were unaware of that fact, we could work out the details,

$$\begin{aligned}
& \int_0^\infty \theta \ p(\theta; \mathbf{y}) \ d\theta \\
= \ & \int_0^\infty \frac{(b+n\bar{y})^{a+n}}{\Gamma[a+n]} \theta^{a+n} \exp[-(b+n\bar{y})\theta] \ d\theta \\
= \ & \frac{(b+n\bar{y})^{a+n}}{\Gamma[a+n]} \int_0^\infty \theta^{a+n+1-1} \exp[-(b+n\bar{y})\theta] \ d\theta \\
= \ & \frac{(b+n\bar{y})^{a+n}}{\Gamma[a+n]} \frac{\Gamma[a+n+1]}{(b+n\bar{y})^{a+n+1}} \int_0^\infty \frac{(b+n\bar{y})^{a+n+1}}{\Gamma[a+n+1]} \theta^{a+n+1-1} \exp[-(b+n\bar{y})\theta] \ d\theta \\
= \ & \frac{(b+n\bar{y})^{a+n}}{\Gamma[a+n]} \frac{\Gamma[a+n+1]}{(b+n\bar{y})^{a+n+1}} \qquad (17) \\
= \ & (a+n)/(b+n\bar{y}). \qquad (18)
\end{aligned}$$

Line (17) follows because the integral in the preceding line equals 1. (Why?) Line (18) uses the fact that $\Gamma[x+1] = x\Gamma[x]$ for any $x \in (0, \infty)$.

**f. Resolution of motivating case studies**

Our first motivating case study asked why we should define

$$df := \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

when testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ using

$$t_{unequal} := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

We are now in a position to answer that question.

Since $(n_1-1)S_1^2/\sigma_1^2$ and $(n_2-1)S_2^2/\sigma_2^2$ are independent chi-square random variables on $(n_1-1)$ and $(n_2-1)$ degrees of freedom, the expected value and variance of

$$S_1^2/n_1 + S_2^2/n_2 = \frac{\sigma_1^2}{n_1(n_1-1)}\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{\sigma_2^2}{n_2(n_2-1)}\frac{(n_2-1)S_2^2}{\sigma_2^2}$$

are

$$\frac{\sigma_1^2}{n_1(n_1-1)}(n_1-1) + \frac{\sigma_2^2}{n_2(n_2-1)}(n_2-1) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{19}$$

and

$$\left(\frac{\sigma_1^2}{n_1(n_1-1)}\right)^2 2(n_1-1) + \left(\frac{\sigma_2^2}{n_2(n_2-1)}\right)^2 2(n_2-1) = \frac{2\sigma_1^4}{n_1^2(n_1-1)} + \frac{2\sigma_2^4}{n_2^2(n_2-1)} \tag{20}$$

respectively.

Suppose that we want to approximate $S_1^2/n_1 + S_2^2/n_2$ by $aY$, where $a$ is a positive real number and $Y$ is a chi-square random variable on $\nu \in (0, \infty)$ degrees of freedom. By analogy to the method of moments, we can compute the mean and variance of $aY$, set them equal to (19) and (20), and then solve for $a$ and $\nu$. We have

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = a\nu \tag{21}$$

and

$$\frac{2\sigma_1^4}{n_1^2(n_1-1)} + \frac{2\sigma_2^4}{n_2^2(n_2-1)} = 2a^2\nu. \tag{22}$$

Dividing (22) by twice (21) yields

$$a = \left( \frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)} \right) / \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right),$$

while dividing twice the square of (21) by (22) yields

$$\nu = \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^2 / \left( \frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)} \right).$$

Thus,

$$S_1^2/n_1 + S_2^2/n_2 \approx \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right) \nu^{-1} Y,$$

so that

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \approx \left( \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) / \sqrt{\nu^{-1} Y},$$

which has the $T$ distribution on $\nu$ degrees of freedom under $H_0 : \mu_1 = \mu_2$. Finally, since $\sigma_1^2$ and $\sigma_2^2$ are unknown, we estimate $\nu$ by

$$df := \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Our second motivating case study asked how to estimate the risk that a randomly selected person from Region B becomes a case in 2010, when there were no cases in Region B during 2009. A Bayesian paradigm will help us here.

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed with probability mass function $f(y; \theta) := \theta^y (1 - \theta)^{1-y}$ for $y \in \{0, 1\}$ and $\theta \in \Theta := [0, 1]$. In the context of our second motivating case study, we can interpret $\theta$ as the risk of a randomly selected individual becoming a case, $Y_k = 1$ as indicating that individual $k$ was a case, and $Y_k = 0$ as indicating that individual $k$ was not a case. We can then describe the observed data from Region B during 2009 as $Y_1 = Y_2 = \cdots = Y_{80{,}000} = 0$. This yields the likelihood function

$$L(\zeta; \mathbf{y}) := (1 - \zeta)^{80{,}000},$$

which is obviously maximized at $\zeta = 0$.

However, we can impose the prior distribution

$$p(\theta) \propto \theta^a (1-\theta)^b,$$

where $a$ and $b$ are nonnegative constants. Since the posterior distribution has the form

$$p(\theta; \mathbf{y}) \propto \theta^a (1-\theta)^{80,000+b},$$

we see that $a$ and $b$ (if integers) can be interpreted as if they came from an auxiliary sample in which there were $a$ cases and $b$ non-cases. Moreover, by solving the equation

$$\frac{\partial\{\theta^a(1-\theta)^{80,000+b}\}}{\partial\theta} = 0,$$

we find that the posterior mode is $a/(a + 80,000 + b)$. This is nonzero as long as $a$ is positive.

The question then becomes, what are sensible choices of $a$ and $b$?

If we took $a$ to be the total number of cases in all regions except B and $b$ to be the total number of non-cases in all regions except B, then the posterior mode would be $20/1,000,000$ or $2.0$ per $100,000$. In effect, we would be collapsing the six regions together to produce an estimate of the risk.

A less drastic option is to regard the observed incidence rates from 2009 ($1/50,000$; $0/80,000$; $6/120,000$; $5/200,000$; $2/250,000$; $6/300,000$) as the realization of a random sample from the beta distribution with parameters $a+1$ and $b+1$ and then apply the method of moments to estimate $a$ and $b$.

The mean and variance of the observed incidence rates are $2.050 \times 10^{-5}$ and $2.935 \times 10^{-10}$ respectively. Solving

$$\frac{a+1}{a+b+2} = 2.050 \times 10^{-5} \quad \text{and} \quad \frac{(a+1)(b+1)}{(a+b+2)^2(a+b+3)} = 2.935 \times 10^{-10}$$

for $a$ and $b$ yields (verify this at home, for practice) $a = 0.432$ and $b = 69,842.818$. We then obtain a posterior mode of $0.432/(149,843.250)$ or $0.3$ per $100,000$. This is a compromise between what maximum likelihood would produce for Region B alone ($0.0$ per $100,000$) versus what it would produce if all six regions were combined ($2.0$ per $100,000$).

Finally, the use of data to specify parameters in the prior distribution is sometimes called "Empirical Bayesian" inference.